How to Build a Cluster

Intel® Server Board S5000PAL

Recipe ID: 15PLTF240000000100-04

Contents

http://www.intel.com/go/esaa

The information contained in this document is provided for informational purposes only and represents the current view of Intel Corporation ("Intel") and its contributors ("Contributors") on, as of the date of publication. Intel and the Contributors make no commitment to update the information contained in this document, and Intel reserves the right to make changes at any time, without notice.

DISCLAIMER. THIS DOCUMENT, IS PROVIDED "AS IS." NEITHER INTEL, NOR THE CONTRIBUTORS MAKE ANY REPRESENTATIONS OF ANY KIND WITH RESPECT TO PRODUCTS REFERENCED HEREIN, WHETHER SUCH PRODUCTS ARE THOSE OF INTEL, THE CONTRIBUTORS, OR THIRD PARTIES. INTEL, AND ITS CONTRIBUTORS EXPRESSLY DISCLAIM ANY AND ALL WARRANTIES, IMPLIED OR EXPRESS, INCLUDING WITHOUT LIMITATION, ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR ANY PARTICULAR PURPOSE, NON-INFRINGEMENT, AND ANY WARRANTY ARISING OUT OF THE INFORMATION CONTAINED HEREIN, INCLUDING WITHOUT LIMITATION, ANY PRODUCTS, SPECIFICATIONS, OR OTHER MATERIALS REFERENCED HEREIN. INTEL, AND ITS CONTRIBUTORS DO NOT WARRANT THAT THIS DOCUMENT IS FREE FROM ERRORS, OR THAT ANY PRODUCTS OR OTHER TECHNOLOGY DEVELOPED IN CONFORMANCE WITH THIS DOCUMENT WILL PERFORM IN THE INTENDED MANNER, OR WILL BE FREE FROM INFRINGEMENT OF THIRD PARTY PROPRIETARY RIGHTS, AND INTEL, AND ITS CONTRIBUTORS DISCLAIM ALL LIABILITY THEREFOR.

INTEL, AND ITS CONTRIBUTORS DO NOT WARRANT THAT ANY PRODUCT REFERENCED HEREIN OR ANY PRODUCT OR TECHNOLOGY DEVELOPED IN RELIANCE UPON THIS DOCUMENT, IN WHOLE OR IN PART, WILL BE SUFFICIENT, ACCURATE, RELIABLE, COMPLETE, FREE FROM DEFECTS OR SAFE FOR ITS INTENDED PURPOSE, AND HEREBY DISCLAIM ALL LIABILITIES THEREFOR. ANY PERSON MAKING, USING OR SELLING SUCH PRODUCT OR TECHNOLOGY DOES SO AT HIS OR HER OWN RISK.

Licenses may be required. Intel, its contributors and others may havepatents or pending patent applications, trademarks, copyrights or other intellectual proprietary rights covering subject matter contained or described in this document. No license, express, implied, by estoppel or otherwise, to any intellectual property rights of Intel or any other party is granted herein. It is your responsibility to seek licenses for such intellectual property rights from Intel and others where appropriate.

Limited License Grant. Intel hereby grants you a limited copyright license to copy this document for your use and internal distribution only. You may not distribute this document externally, in whole or in part, to any other person or entity.

LIMITED LIABILITY. IN NO EVENT SHALL INTEL, OR ITS CONTRIBUTORS HAVE ANY LIABILITY TO YOU OR TO ANY OTHER THIRD PARTY, FOR ANY LOST PROFITS, LOST DATA, LOSS OF USE OR COSTS OF PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES, OR FOR ANY DIRECT, INDIRECT, SPECIAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF YOUR USE OF THIS DOCUMENT OR RELIANCE UPON THE INFORMATION CONTAINED HEREIN, UNDER ANY CAUSE OF ACTION OR THEORY OF LIABILITY, AND IRRESPECTIVE OF WHETHER INTEL, OR ANY CONTRIBUTOR HAS ADVANCE NOTICE OF THE POSSIBILITY OF SUCH DAMAGES. THESE LIMITATIONS SHALL APPLY NOTWITHSTANDING THE FAILURE OF THE ESSENTIAL PURPOSE OF ANY LIMITED REMEDY.

Intel, the Intel logo, and Intel Xeon are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

*Other names and brands may be claimed as the property of others.

Copyright © 2007, Intel Corporation. All Rights Reserved.

Introduction

This Intel ESAA recipe describes the steps required to build a High Performance Computing (HPC) cluster based upon the Intel S5000PAL server board; from hardware assembly through performance testing and cluster application development. It includes various options for hardware design, interconnects, management, BIOS and OS provisioning, job schedulers, cluster tools and cluster application development packages and runtime environments.

Note: It is recommended that the entire document is read prior to any installation activity

Overview

As mentioned above, this recipe illustrates the beginning to end setup and implementation of a HPC cluster. The basic outline of the recipe is as follows:

1) Cluster hardware setup – racks, power, cabling, switches, etc.

2) Server management – application options for pre boot and post boot system management

3) Node BIOS provisioning – a PXE environment to boot nodes to DOS for updates

4) Node OS provisioning and Cluster Management with Platform Open Cluster Stack (OCS) – An OS provisioning solution for easily deploying, running and managing cluster operating environments

5) Cluster Application Development and Application Execution Environments – Complete packages for running, developing and optimizing cluster applications, and testing cluster performance.

Each of these sections will include either references to locations of existing documentation or detailed instructions on how to setup specific installation requirements.

Hardware Components

Quantity per node	Item	Manufacturer	Model
1	Intel® Server Board	Intel	S5000PAL
1	Intel® Server Chassis	Intel	SR1500
1	Intel® HDD Backplane – SR1500 1U SATA/SAS passive	Intel	ASR1500PASBP
4GB Per Server Board	Memory	Any supported	Please refer to the Tested Memory List at http://www.intel.com/support/motherboards/server/s5000pal/sb/CS-022919.htm
2 Per Server Board	Intel® Xeon® Processors	Intel	Please refer to the Qualified and Supported Processor List at Please refer to the Tested Memory List at http://www.intel.com/support/motherboards/server/sb/CS-022346.htm
2 Per Server Board	1U passive heatsink (light weight with heat pipe)	Intel	AXXUPHS
1 (minimum) Per Server Board	SATA 3.5" hard drive	Any supported	Please refer to the Intel® Tested Hardware and Operating System List at http://www.intel.com/support/motherboards/server/sb/CS-022920.htm
1 Per Server Board	PCI-E InfiniBand* Host Channel Adapter	QLogic	7104-HCA-LPX2P - Dual port PCIExpress 8x SDR MemFree 7104-HCA-LPX2P-DDR - Dual port PCIExpress 8x DDR MemFree Please refer to SilverStorm (QLogic) recipes for cluster configurations from 8 to 256 nodes for SDR or DDR
1 Optional Per Server Board	InfiniBand* Host Channel Adapter	Qlogic	QLE 7140-CK - Single Port 10 GBs InfiniBand to x8 PCI Express Adapter Please refer to SilverStorm (QLogic) recipes for cluster configurations from 8 to 256 nodes for SDR or DDR

1 Optional Per Server Board	InfiniBand* Host Channel Expansion Module	Intel	AXXIBIOMOD
As Needed	InfiniBand cables	Any supported	
1 (min per cluster)	InfiniBand switch	Any supported	
1 (min), As Needed	Ethernet cables	Any supported	
1 (min per cluster)	Ethernet switch	Any supported	
1 (min per cluster)	Keyboard, Video, Mouse (KVM) switch	Any suppo rted	
1	KVM cables	Any supported	

Table 1 – Hardware Bill of Materials

Software Used in the Installation

Dist. By	Description	File Name
Platform Computing	Open Cluster Stack (OCS) 4.1.1-1.1 (RHEL 4 Update 3) or Open Cluster Stack (OCS) 4.1.1-2.1 (RHEL 4 Update 4)	
Intel	Intel® Tools 4.1.9.	
Intel	System Management Software	

Table 2 - Software Bill of Materials

Hardware Installation

Server racks

Depending on the size of the cluster being built, the Front End, switches and any management nodes should be centrally located if multiple racks are being installed. Cluster nodes should follow a clear and organized numbering scheme to aid in node identification. Cable, power, and cooling plans are very important and must be carefully considered. These plans drive the ultimate layout of equipment in the racks. A typical configuration involves switches and servers in the same rack. An example of a large rack layout is illustrated by figure 1 below.

Note: Platform compute node numbering begins at 0. Rack identification numbers begin at 1.*

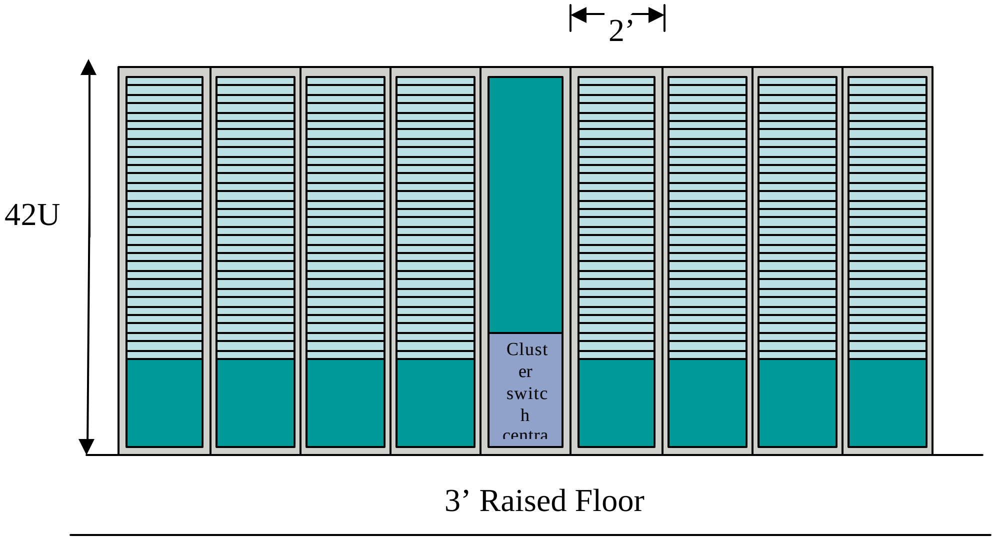


Figure 1. Server racks

Remote console configuration / KVMIP

Using a KVMIP solution is one method for supplying remote console access to all cluster nodes. Connect and configure the supported KVMIP solution such that there is remote console access for each node in the cluster.

Ethernet Switches and Cables

Aside from the Front End having more than one Ethernet interface, cluster nodes should only require a single Ethernet interface for management communication. Ethernet switches and cables used should be Gigabit capable. At least one Ethernet switch should be used per cluster. Larger clusters may require higher capacity switches or tiered switch configurations.

If "spanning tree" is enabled on the switch, it dramatically slows PXE installation because each port in the switch is determining where it fits in the spanning tree to avoid loops in the network. Use caution when changing the spanning tree configuration options on the switch.

Infiniband Switches and Host Channel Adapters

Within each server, a host channel adapter (HCA) must be installed. When installing an HCA, consult either the server motherboard documentation or server vendor to ensure that the fastest available PCI-E slot is selected.

> *Note: The HCA should be the only device on the bus.*

If they are used, it is also recommended that you place any InfiniBand switch(s) at the bottom of a rack. This allows any inter-rack cables to be cleanly routed below the floor.

Handling InfiniBand Cables

InfiniBand cables are more sensitive than other types of networking cables (such as Ethernet and Fiber Channel); therefore, greater care must be taken when handling them.

Rotating InfiniBand Connectors

If the InfiniBand cable connector is not properly oriented to fit onto the port receptacle when you attempt to insert it, do not twist the connector to achieve the correct orientation.

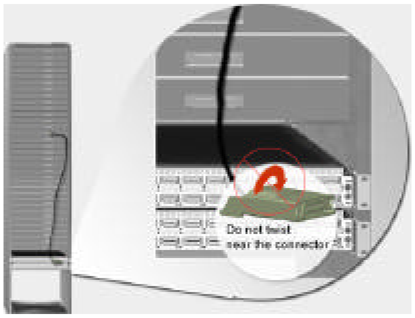


Figure 2. Rotating InfiniBand* Connectors

Instead, reach back a few feet on the cable, and twist the bulk cable to allow the connector to rotate to the proper orientation. Doing this prevents applying all of the rotational force directly at the plug terminations.

Minimum Cable Bend Radius

InfiniBand cables can be damaged if they are bent beyond the minimum bend radius. Damage can occur even if the cables are temporarily bent beyond this limit during installation. It is helpful to "pre-form" the cables to the bent condition prior to installation. This reduces any undue strain at the plug or receptacle connection at the time of installation. Figure 3 illustrates the areas that are of particular concern, and Figure 4 provides measurement guidelines.

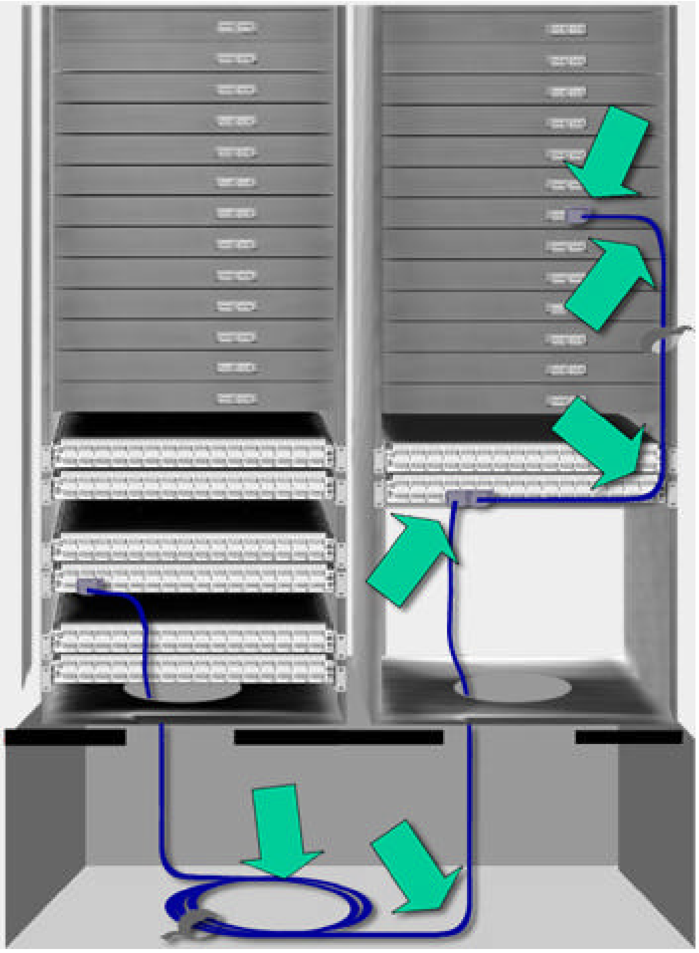
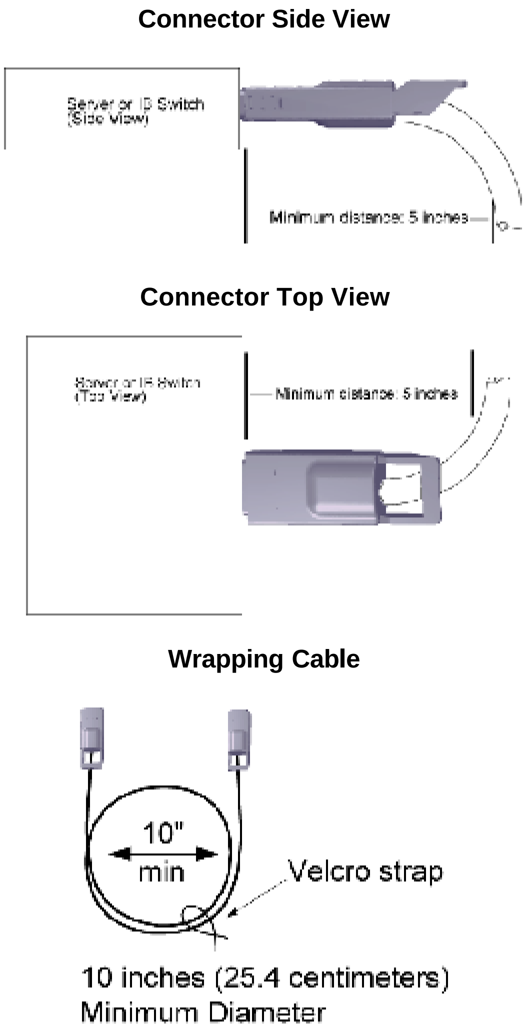


Figure 3. InfiniBand* Cable Bend Radius



Connector Side View

Connector Top View

Wrapping Cable

Figure 4. InfiniBand* Cable Bend Radius Recommendations

Server Management

The S5000PAL server board comes with a variety of management utilities for updates and maintenance. The Intel System Management Software package will enable remote maintenance and management of the cluster nodes from a centralized system. Download the package here:

http://downloadcenter.intel.com/scripts-df-external/filter_results.aspx?strTypes=all&ProductID=2363&OSFullName=All+Operating+Systems&lang=eng&strOSs=All&submit=Go%21

The Intel System Management Software provides a lower level of management to the user for remote operations and maintenance. Download the package here:

http://downloadcenter.intel.com/scripts-df-external/Product_Filter.aspx?ProductID=2363

BIOS and Firmware Provisioning

At times it may become necessary to update some or all nodes to the latest revisions of BIOS and Firmware and set all the required BIOS settings for optimal performance. This section explains how this is accomplished in a multi node compute cluster using existing recipes available from the Intel ESAA web site. Once you have assembled your cluster and are ready to perform any or all BIOS and Firmware provisioning, go to the following website and download the recipe "BIOS and Firmware Provisioning for Intel© Server Board S5000PAL."

http://www.esaa-members.com

This recipe illustrates in detail how to build a PXE boot environment on a Red Hat RHEL 4 U3 server. This will allow the cluster nodes to boot up and automatically update the necessary BIOS depending on what is required. It is a fully configurable custom DOS boot image. It is assumed that the user will have a working understanding of DOS and how to make, edit and automate batch files and executables. For the Intel® Server Board S5000PAL, go to the website below to download the most current BIOS update package:

http://downloadcenter.intel.com/scripts-df-external/filter_results.aspx?strTypes=all&ProductID=2451&OSFullName=All+Operating+Systems&lang=eng&strOSs=All&submit=Go%21

Unzip these files and place them into the DOS image created from the provisioning recipe above. Modifying some of the batch files may be needed. To execute multiple batch files from within a bootable DOS image, modify the autoexec.bat to call other batch files. An example autoexec.bat file is included in the update download.

Once each node has been booted to the DOS image it will go through the entire update as specified in the autoexec.bat file. Once finished, it will wait for a prompt from the user. At this point each node will be ready to be rebooted manually, via Intel® AMT or with Intel® System Management.

> *Note: Make sure that the BIOS and Firmware provisioning PXE server is shutdown prior to rebooting the nodes or they will go into a reboot loop and provision the BIOS and Firmware again.*

At this point the cluster is ready for Operating System (OS) provisioning. Ensure that the provisioning solution begins here so that the next time the cluster nodes are rebooted they will boot from the OS provisioning server.

OS Provisioning and Cluster Management

The cluster OS provisioning method deployed is up to the user(s). The ESAA solutions below offer several options on OS provisioning with various high speed interconnects.

Platform OCS 4.1.1-1.1/4.1.1-2.1

The Platform Computing OS provisioning solution offers an end to end solution for deploying clusters from a Front End including options for either Intel or SilverStorm high speed interconnect support. The Platform Computing OCS 4.1.1-1.1/4.1.1-2.1 ("High-Performance Computing (HPC) Cluster Installation using Platform Open Cluster Stack (OCS)* 4.1 on Red Hat* Enterprise Linux* 4.3 - Intel® Server Board S5000PAL" or "Platform Open Cluster Stack (OCS)* 4.1.1- 2.1 on Red Hat Enterprise Linux* 4.4 - Intel® Server Board S5000PAL") recipes for this platform can be found on the ESAA website:
http://www.esaa-members.com
Once you have a Linux OS installed, and the IB interconnect has been installed with the SilverStorm RPM, the separate SilverStorm recipe for (High-Performance Computing (HPC) Cluster Installations using SilverStorm* InfiniBand* Interconnect on Linux* - Intel® Server Board S5000PAL) can be used as reference for additional installation and configuration. (http://www.esaa-members.com)

Z RESEARCH GlusterHPC

GlusterHPC automates installation of High Performance Computational Clusters / Supercomputers on commodity Intel platform. It is extensible, portable across various GNU/Linux distributions, scalable to several thousand nodes and has clean easy to use dialog interface. GlusterHPC provides HPC tools and libraries in addition to OS packages and pre-configures them at the time of provisioning. It essentially converts a stand-alone GNU/Linux distribution into a Cluster distribution.

The Z RESEARCH* GlusterHPC* Installation - Intel® Server Board S3000AH recipe for this platform can be found on the ESAA website:

http://www.esaa-members.com

Cluster Tools

Once the cluster has been deployed, there are several tool suites that can be used to verify functionality and tune performance. Included in the Platform* OCS* installation is the Intel® MPI Library 2.0.1runtime environment. Additionally, the Intel® Tools suite version 4.1.9 can be installed during the Platform OCS 4.1.1-1.1/4.1.1-2.1 install as well as post install." It may be downloaded from the link below:

http://my.platform.com

To install the Intel® Tools 4.1.9. Roll, use the Platform* OCS* installation instructions detailing how to install additional rolls. Intel® Tools suite 4.1.9 includes Intel Compilers, however to compile applications installing licenses is required. (If the licenses are not installed, the compiling functionality of the Intel Compilers will not be available; however a runtime environment will still be available.)

The licenses can be installed with the following steps. Perform the following as root prior to rolling any compute nodes:

1) Copy license files to be used on the Front End to: /opt/intel/licenses

2) Create the directory: /export/apps/intel/licenses

3) Copy license files to be used on the compute nodes to: /export/apps/intel/licenses

4) Create a script called intel_license.sh. The script should be:

```
#!/bin/sh

rm -rf /opt/intel/licenses

ln -s /share/apps/intel/licenses /opt/intel/licenses
```

5) Use rocks-compute to add the script to the post-install

```
# rocks-compute -s intel_license.sh -b
```

6) New licenses can be added by placing them in: /export/apps/intel/licenses on the Front End.

If this optional roll was not installed, another choice is to install the Intel Cluster Toolkit for Linux. It can be obtained at the following URL:

http://www3.intel.com/cd/software/products/asmo-na/eng/cluster/244171.htm